COLLISION ENERGY SPECIFIC TANDEM MASS SPECTRAL PREDICTION

Technical Field

The present invention relates, in general terms, to systems and methods for predicting collision energy specific tandem mass spectra for molecular fingerprinting.

Background

The reference in this specification to any prior publication (or information derived from it), or to any matter which is known, is not, and should not be taken as an acknowledgment or admission or any form of suggestion that that prior publication (or information derived from it) or known matter forms part of the common general knowledge in the field of endeavour to which this specification relates.

Liquid chromatography tandem mass spectrometry (LC-MS/MS) is a crucial analytical tool for unravelling complex mixtures and identifying compounds with applications in food, agrochemicals, diagnostics, environmental contaminants, and therapeutics amongst others. The ability to identify individual compounds within a sample mixture quickly and accurately is often critical for further research, development, or testing. Typically, this is achieved by matching their tandem mass spectral (MS/MS) fingerprint against a reference MS/MS library of known compounds. However, current spectral reference libraries are incomplete or limited. Furthermore, MS/MS spectra of identical molecules vary significantly based on the collision energy applied during analysis, resulting in the need for multiple reference spectra for each molecule.

For example, in PubChem, the world's largest chemical database, the estimated number of compounds with corresponding MS/MS spectra is <1%. This limited number of reference MS/MS spectra makes it difficult to quickly identify

compounds. Moreover, expansion of MS/MS reference libraries through experimental means is both costly (SGD$440 / sample, https://www.ntu.edu.sg/brc/research-capabilities/specialised-facilities/mass-spectrometry) and time-consuming.

It would be desirable to provide a method for predicting collision energy specific tandem mass spectra for specific molecules.

Summary

In view of the limitations in prior art methodologies for mass spectral fingerprinting of compounds, the methodology described herein generates collision energy specific tandem mass spectra from corresponding molecular structures with ~75% accuracy to experimental spectra. This enables structural identification of molecules directly from MS/MS spectra.

As purification, isolation, and subsequent structural identification often bottlenecks progress, direct structural identification from MS/MS can significantly accelerate efforts in molecular discovery (e.g. in untargeted metabolomics, natural product discovery) or in high throughput mass spectrometry analysis workflows (e.g. identifying biomarkers, odorants, metabolites in mixtures).

Disclosed is a method for training a model to predict collision energy specific tandem mass spectrum (MS/MS) of a molecule, comprising:

receiving input data comprising molecules and, for each molecule, an adduct type, a molecular formula, Simplified Molecular Input Line Entry System (SMILES) string, chemical identifier for the molecule, and ground truth intensities of a mass spectrum generated for the molecule at a specific collision energy;

generating a directed acyclic graph (DAG) comprising predicted fragments for each molecule based on the respective adduct type,

SMILES string, chemical identifier, and the ground truth intensities;

training a fragmentation model based on the DAG for each molecule and the respective molecular input and ground truth intensities, to generate most probable fragments based on molecular inputs;

separating the input data into bins corresponding to non-overlapping collision energy ranges for which identical ones of the molecules have mass spectra with at least a minimum threshold similarity;

freezing the fragmentation model and generating further fragments, using the frozen fragmentation model, from input data corresponding to further molecules and respective ground truth mass spectra each generated at a specific collision energy, and training a plurality of intensity models, each intensity model being trained on the input data corresponding to a respective one of the bins, to predict a spectrum each of the further molecules, by:

predicting an intensity of a MS/MS peak for a mass of each further fragment and for a range of masses corresponding to potential hydrogen additions to, and hydrogen removals from, the further fragment; and

updating the respective intensity model to maximise weighted cosine similarity between the predicted MS/MS peak intensities and the ground truth mass spectrum.

Also disclosed is a model trained in accordance with the method set out above.

Also disclosed is a tandem mass spectral reference library comprising:

a plurality of molecules and, for each molecule, one or more mass spectra comprising peak intensities predicted using the model described above.

Brief description of the drawings

Embodiments of the present invention will now be described, by way of non-

limiting example, with reference to the drawings in which:

Figure 1A provides an overview of technology to generate a collision energy specific tandem mass spectral reference library, and Figure 1B provides a more detailed workflow for model 100 of Figure 1A.

Figure 2 shows weighted cosine similarity comparisons of the tandem mass spectra of example molecule DL-5-hydroxylysine across different collision energy values.

Figure 3, comprising images (A) and (B), shows results of transfer learning on weighted cosine similarity, with image (A) showing the distribution of weighted cosine similarity scores comparing generated and ground truth spectra, and image (B) showing the distribution of Recall@X scores.

Figure 4, comprising images (A) and (B), provides an overview of Recall@X calculation in which image (A) replaces ground truth MS/MS spectra of the compound with MS/MS spectra generated by the model 100 of Figure 1, and image (B) ranks search querying the ground truth MS/MS spectra against the library using weighted cosine similarity (CS = weighted cosine similarity).

Figure 5 shows transfer learning model performance on a 69/71 eV validation dataset with varying amounts of training data.

Figure 6 compares weighted cosine similarity between tandem mass spectra (MS/MS) of Pterin at 5 different collision energies (10 eV, 25 eV, 40 eV, 50 eV, 70 eV).

Detailed description

A collision energy-specific machine learning model that can quickly (<1 second/molecule) and accurately (75% average cosine similarity) predict

MS/MS fingerprints from molecular structures. This can be used to generate a high-fidelity MS/MS spectral reference library faster and at much lower cost compared to experimental means.

Figure 1A provides an overview of the various contributions to mass spectral fingerprinting taught herein. In particular, Figure 1 demonstrates a model 100 for predicting collision energy specific mass spectra for molecules (i.e., a collision energy specific molecule-to-MS/MS model), molecule generation model 200 for expanding a library of molecules for which mass spectra can be predicted, and a tandem mass spectral library 300 generated using generated molecules (from models 100 and 200).

Collision energy specific molecule-to-MS/MS model 100

Model 100 flexibly caters to specific collision energies, using fragmentation model 102 and intensity models 104. The fragmentation model 102 generates fragments based on a molecular structure. Transfer learning has been used to fine-tune the intensity models 104 to predict collision energy specific intensities for generated fragments.

The model 100 balances data requirements against the accuracy of the predicted mass spectra. This is achieved by dividing the range of collision energies over which mass spectra are produced, into a plurality of bins. The bins cover respective non-overlapping collision energy ranges for which mass spectra for a particular molecule can be expected to differ by at most a predetermined amount. This predetermined amount corresponds to a minimum similarity threshold. For example, the minimum similarity threshold may be 95%, 94% or some other amount, being the cosine similarity between mass spectra, for the same molecule, for a range of collision energies. The model 100 has two stages. the first stage uses a model 102 to generate or predict fragments of an input molecule. A single model 102 predicts fragments, regardless of the number of bins. The trade-off between size and accuracy is in the second stage – i.e., in

training multiple intensity models 104 for predicting MS/MS spectra, where there is one model 104 for each bin.

The method 100 itself is used for training a model to predicting a collision-energy specific tandem mass spectrum (MS/MS) for a molecule. The method 100 involves receiving input data (step 106) comprising a molecule and ground truth mass/charge (m/z) peaks (intensities) of a mass spectrum generated at a specific collision energy, and generating a directed acyclic graph (DAG – step 108) that represents fragmentation pathways of the molecule – i.e., the different fragments into which the molecule can be broken. The input data and DAG are then used to train a fragmentation model 102 (per step 110). The fragmentation model thus learns to generate fragments based on a molecular input – e.g., the SMILES or InChiKey corresponding to the molecule.

The input data, specifically the ground truth m/z peaks at specific collision energies are then separated into bins (step 112). The fragmentation model is then frozen (step 114 – though the fragmentation model can be frozen before step 112) and the intensity models 104 are trained. Training the intensity model involves inputting adduct types, molecular inputs (e.g., SMILES or InChiKey) of molecules, and ground truth intensity peaks for those molecules. The trained, frozen fragmentation model 102 generates fragments for each molecule and the intensity model 104 predicts m/z peaks corresponding to the fragments (step 116). The predicted peaks are compared to the ground truth peaks using a weighted cosine similarity (step 118) to be maximised during model training. The intensity models 104 are then trained and outputted.

During training, the model 100 receives input data (step 106). The input data comprises molecules and, for each molecule, an adduct type, the molecular formula, the Simplified Molecular Input Line Entry System (SMILES) string, a chemical identifier for the molecule (e.g., the InChIKey), and ground truth intensities of a mass spectrum generated for the molecule at a specific collision energy.

Per step 108, a directed acyclic graph (DAG) is then generated for each molecule based on the respective adduct type, SMILES string, chemical identifier, and the ground truth intensities. The DAG is trained to recursively select the atoms in an input molecule, around which to break bonds. Predictions are made for the input molecule, regarding the atoms around which to break the bonds, to generate the fragments. The substructures (fragments) are then chemiformatically generated, and further bond breakages are iteratively applied.

To train this model, a "ground truth DAG" is used, as determined by an existing combinatorial enumeration algorithm or exhaustive DAG generation model such as MAGMa. The DAG is generated from the adduct type, SMILES string, chemical identifier, and the ground truth intensities in the input data. The DAG represents fragmentation pathways of the molecule, and thus predicts fragments that will be generated.

The root fragment of the DAG is the input molecule. The predicted fragmentation probability for the bonds around each atom in each fragment can be cheminformatically estimated. Predictions are mode atom-wise (i.e., at every single atom for each fragment in the DAG). To make this atom-wise prediction, information about the root molecule, fragment molecule, their difference, their respective chemical formulas, the adduct, and the number of bonds that were broken between the root molecule and fragment are encoded. To embed the root molecule, a gated graph neural network (GNN) is used, where embeddings are pooled across atoms (specified by a hyperparameter). The gated GNN uses the same network to learn representations of the fragment. For all graph neural networks, a one-hot encoding of the adduct type is also added as atom-wise features alongside the bond types and atom types. A one hot vector is also generated for the number of bonds broken. This can then be used to shift spectra in the intensity model 104, based on the mass of Hydrogen atoms added or subtracted from a fragment.

The model is trained to maximize the probability of generating the DAG by minimizing the binary cross entropy loss over each atom for every fragment in an observed spectrum.

The fragmentation model 102 is then trained (step 110) using the DAG generated at step 108 for each molecule, the respective molecular formula and ground truth intensities. The fragmentation model 102 generates the most probable fragments based on molecular formulae and adduct type. For input data from which the fragmentation model 102 generates fragments, the fragments correspond to (and are thus annotated with) ground truth intensities generated at a specific collision energy.

The ground truth data was generated during actual LC-MS/MS at the specific collision energy. During experiments, the input data was taken from the National Institute of Standards and Technology (NIST) MS/MS Library 2020 (NIST20) – ending up with 246,581 spectra of 18,384 unique compounds at varying collision energies. Optimization of transfer learning parameters was carried out on the subset of data encompassing the MS/MS spectra of 1,159 compounds at 70 eV (split into 56% Train/19% Validate/25% Test) and then further validated against the subset of 2,364 MS/MS spectra taken at 69 eV and 71 eV (see Annex B for transfer learning details).

The performance of machine learning models rely heavily on the quality and quantity of available training data. The minimum amount of data required for optimal transfer learning performance was determined based on where model accuracy plateaued, as shown in Figure 5. In particular, Figure 5 shows that a minimum of 113 datapoints (5% training data) is required for optimal training of binned transfer learning models, with more data giving diminishing returns as model performance plateaus.

Collision energy bins for training model 104

Once trained, the fragmentation model is frozen (step 114). It can then be used to generate fragments based on further molecules. The input data is separated into bins (step 112 – which can occur before or after step 114) corresponding to non-overlapping collision energy ranges for which identical ones of the molecules have mass spectra with at least a minimum threshold similarity. An intensity model 104 can then be trained for each bin, using only that data which corresponds to the respective bin – i.e., a separate intensity model will be trained for each bin, using data generated at a specific collision energy that falls within the range corresponding to the respective bin.

In some embodiments, the ranges of collision energies for the bins are equal – for example, 10 eV (0-9 eV, 10-19 eV, 20-29 eV etc) – this can be set by averaging the weighted cosine similarity across all bins (i.e., some bins will have similarity greater than the threshold where other bins will have similarity slightly lower than the threshold) or by reducing the collision energy range of the bins until the lowest similarity is equal or greater than the threshold similarity. In other embodiments, the ranges are determined by starting at one end of the spectrum are comparing the cosine similarity between spectra for progressively different collision energies until the variation exceeds the minimum similarity threshold, each bin being defined by the collision energy at which the last spectrum was generated that met the minimum similarity threshold. For example, starting at 0 eV collision energy, mass spectra for the same molecule may be compared at 1 eV, 2 eV and so on, with all mass spectra meeting the minimum similarity threshold until, and including, 8 eV. If the mass spectrum at 9 eV does not meet the minimum similarity threshold, then the bin covers the range of 0 eV to 8 eV, and the process is repeated for the next bin, starting at 9 eV.

MS/MS fragmentation patterns of identical molecules change significantly with different collision energies applied. Ideally, each collision energy should have its own model. Training a new model comes at a high computational cost and

associated data storage requirements. To avoid generating a specific model for every collision energy, a model 104 is trained for each bin of 10 eV – for a collision energy range of 0 eV to 129 eV, there may be 13 non-overlapping bins. Each bin includes input data generated over a specific range of collision energies. The specific ranges are generally non-overlapping so that each input data point appears in one bin only. To determine the range of collision energies for any bin, an assessment was made to ensure all mass spectra for experiments conducted over a range of collision energies had a minimum threshold similarity – e.g., weighted cosine similarity.

In experiments, a collision energy method will be employed to adjust the collision energy applied based on the molecular weight and charge. For a typical setting of scan range of up to 1,800 m/z and 35% normalized collision energy this works out to about 130 eV. Statistical analysis across 19,024 examples in the NIST20 dataset revealed that weighted cosine similarities of MS/MS spectra from identical molecules were on average 94.8% similar when within ±5 eV collision energy of each other as shown in Figure 2. If the threshold similarity is set to 94.8%, the non-overlapping collision energy ranges will therefore be 10 eV as shown in Table 1.

Bin	eV Range	#Datapoints (NIST20)	Test loss
1	0 – 9	9,425	0.1397
2	10 – 19	15,820	0.2041
3	20 – 29	17,176	0.2661
4	30 – 39	16,984	0.3000
5	40 – 49	16,311	0.3204
6	50 – 59	14,464	0.3194
7	60 – 69	11,826	0.3293
8	70 – 79	9,350	0.3389
9	80 – 89	7,126	0.3437
10	90 – 99	4,966	0.3622
11	100 – 109	3,073	0.3908
12	110 – 119	2,025	0.3891
13	120 – 129	1,398	0.4448

Table 1 – example data from applying collision energy specific molecule-to-

MS/MS model to 13 collision energy bins from 0 – 130 eV.

Consequently, a model finetuned on MS/MS spectra in a specific bin of 10 eV should provide good predictions for MS/MS spectra within the corresponding range of collision energies. As a result, a model 104 is trained for each bin. Further extension of the collision energy bin range would compromise model accuracy while minimizing the range would significantly increase computational costs. In Table 1, a 130 eV range was split into 13 bins of 10 eV width and transfer learning was performed using an 80% train / 10% validate / 10% test split with Version 65 transfer learning parameters – results shown in Figure 3.

With further reference to Figure 1B, the intensity model 104 for each bin is then trained using data for the corresponding bin. During training, fragments are generated by the fragmentation model 102 for each input molecule. For all molecules for which a ground truth mass spectrum (i.e., ground truth intensity peaks) was generated at a collision energy corresponding to a particular bin, the corresponding intensity model 104 trains on the generated fragments and collision energy to predict intensities of peaks of the mass spectrum for the molecule at the specific collision energy. The intensity model 104 also predicts intensities of peaks for a range of masses corresponding to potential hydrogen additions to, and/or removals from, each generated fragment based on the number of bonds broken to generate the fragment. The intensity model 104 for each bin is then updated to maximise the weighted cosine similarity between the predicted MS/MS peak intensities (i.e., the artificial mass spectrum) and the ground truth mass spectrum (step 118). Thus, the weighted cosine similarity is calculated between the predicted MS/MS peak intensities from the intensity model 104, and the ground truth intensity peaks, and is fed back into the intensity model 104 to adjust the weights of the intensity model 104.

The trained model can then be outputted (step 120) for use in predicting m/z peak intensities for new molecules, including artificially generated molecules.

Figure 3, image (A) shows the distribution of cosine similarity scores for generated spectra compared to ground truth spectra. This shows over 75% accuracy compared with a current benchmark of 59% (ORIGINAL = prediction made using the ICEBERG model). Figure 3, image (B) shows the distribution of Recall@X scores, clearly illustrating the efficacy of fingerprinting using the binned protocol where, when compared with experimentally derived spectra in the reference library, each artificial spectrum (i.e., spectrum generated using model 100) is within the top 6.1 spectra on average, based on weighted cosine similarity.

Similarity can be determined on any number of bases. For present purposes, weighted cosine similarity is used. Cosine similarity is a measure of how similar two MS/MS spectra are to each other. Weighted cosine similarity is calculated by:

$$(I_q, I_i) = \frac{\sum_{k=1}^{M_{max}} m_k \sqrt{I_{qk}} \cdot m_k \sqrt{I_{ik}}}{\sqrt{\sum_{k=1}^{M_q} (m_k \sqrt{I_{qk}})^2} \sqrt{\sum_{k=1}^{M_i} (m_k \sqrt{I_{ik}})^2}} \quad (1)$$

where I_q and I_i are vectors of *m/z* intensities representing two spectra, m_k and I_k are the *m/z* ratio and intensity found at *m/z* = k, M_q and M_i are the largest indices of I_q and I_i with non-zero values, and M_{max} is the larger of M_q and M_i. The cosine similarity is weighted by *m/z* as the peaks in mass spectra corresponding to larger fragments are more characteristic and useful in practice for identifying. Figure 6 compares weighted cosine similarity between tandem mass spectra (MS/MS) of Pterin at 5 different collision energies (10 eV, 25 eV, 40 eV, 50 eV, 70 eV).

Recall@X is a measure representing how well generated spectra will perform as a reference in a spectral library. The ground truth MS/MS spectra in the 2,364 spectral library of 69 / 71 eV spectra is first replaced by its generated version by model 100. The ground truth MS/MS spectra is then compared against the entire library via weighted cosine similarity and the library is rearranged in

descending order of weighted cosine similarity score and ranked. The rank at which the generated MS/MS comes in is the Recall@X score shown in Figure 4.

As discussed above, the intensity model predicts a spectrum each of the further molecules, by predicting the intensity of a mass-to-charge (m/z) peak for a mass of each further fragment and maximising the weighted cosine similarity between the predicted MS/MS peaks and the ground truth mass spectrum. The intensity model may also predict m/z peaks for a range of masses corresponding to potential hydrogen additions to, and hydrogen removals from, each fragment. This accounts for variations in bonding structures for isomorphisms. In maximising the cosine similarity, mass spectra generated using the model 100 are compared with the ground truth mass spectra, and the intensity model is iteratively updated to increase weighted cosine similarity.

By freezing the fragmentation model during training of the intensity model, embeddings learned during fragmentation are transferred to the intensity model in the form of the fragments generated by the fragmentation model. Therefore, the intensity model is matched to the fragmentation model, as opposed to training the intensity model on the ground truth fragments. In the latter case, the intensity model and fragmentation model could mismatch, where small errors in the predictions from each model compound. In contrast, in the former case, errors in the fragmentation model can be accounted for by the intensity model.

Molecule generation

In order to generate a comprehensive reference MS/MS library for natural product discovery workflows, 403,819 compounds from the Collection of Open Natural Products (COCONUT, https://coconut.naturalproducts.net/) that are representative of all known natural products to-date, were combined with 67,989,628 natural product-like compounds generated in-house (Tay, D.W.P., Yeo, N.Z.X., Adaikkappan, K. et al. 67 million natural product-like compound

database generated via molecular language processing. Sci Data 10, 296 (2023). https://doi.org/10.1038/s41597-023-02207-x) to give a total of 68,393,447 unique compounds.

To generate the molecule, an LSTM model was trained on tokenized SMILES (with stereochemistry removed) from 325,535 (80%) out of the 406,919 known natural products in COCONUT, the collection of open natural products. The model was able to break down SMILES into unique tokens (e.g. C, N, S, O, c, n, 1, 2..etc), learn how to assemble these token together according to the molecular language of natural products, and generate 100 million natural product-like SMILES with no specified stereochemistry. Although stereochemistry in natural products can confer specific bioactivity, the present pipeline removes stereochemistry to reduce data complexity, lower file size, and improve fidelity of the generated structural database. In any case, a range of feasible stereoisomers for each molecule can still be obtained through iterative enumeration of its 3D structures followed by back transformation to stereospecific SMILES. Following this approach, extended isomer libraries of shortlisted SMILES of interest can be generated to cover wider isomeric space than a database of pre-defined stereospecific SMILES.

Although alternative approaches for the generation of natural product virtual libraries have been attempted, prior libraries have been limited in terms of novelty (frequent re-occurrence of well-known scaffolds), natural product-likeness (43% meeting threshold compared to 85% in the training set), and scale (<1.5 million molecules). Using model 200, an openly available virtual library of >67 million natural product-like SMILES is created, with a distribution of natural product-likeness scores similar to that of known natural products yet encompassing expanded physiochemical and structural space, indicating its potential for *in silico* discovery of natural products.

Cheminformatics toolkits RDKit, ChEMBL chemical curation pipeline, NP Score, and NPClassifier were employed to sanitize, analyze, and characterize the

generated 100 million natural product-like SMILES database.

First, RDKit function Chem.MolFromSmiles() was used to filter out 9,596,585 syntactically invalid SMILES from the 100 million generated set. Second, to ensure molecular uniqueness within the dataset, RDKit functions Chem.MolToSmiles(Chem.MolFromSmiles()) and Chem.inchi.MolToInchi() was used to convert the generated SMILES into canonical SMILES and International Chemical Identifier (InChI) representations for comparison and filtering, resulting in the removal of 22,484,883 (22%) duplicates. Third, the ChEMBL chemical curation pipeline was applied for further sanitization and standardization by checking and validating chemical structures; assigning an error score if structural issues are detected (where error scores increase with the severity of the problem); standardizing chemical structures based on FDA/IUPAC guidelines; and generating parent structures by removing isotopes, solvents, and salts.

Through this process, a further 854,328 invalid molecules with penalty scores exceeding 5 (indicating severe structural issues), were filtered out. Combined with the earlier detected syntactically invalid SMILES, a total of 10,450,913 (11%) invalid generated SMILES were identified and removed. The top 2 structural errors reported amongst the remaining valid molecules were (1) undefined stereochemistry (95%), which was due to the generation of SMILES without stereochemistry, and (2) the need for (de)protonation (2%), which was addressed later in Step 3 of the ChEMBL chemical curation pipeline. On the whole, these pre-processing steps refined the initial dataset down to the reported 67,064,204 valid, unique, natural product-like SMILES generated database.

Fourth, RDKit was used to calculate natural product-likeness scores (NP Score) for both known natural product SMILES and generated SMILES. NP Score employs atom-centred fragments (HOSE codes) and bonding information to characterize structural features and calculate a Bayesian measure of molecular

similarity to known natural product structural space. The NP Score distribution of the generated natural product-like SMILES was found to closely resemble that of known natural products from the COCONUT database with a Kullback-Leibler (KL) divergence of 0.064 nats, supporting that natural product-like molecules had been generated.

Fifth, the NPClassifier toolkit was used to classify both natural product-like SMILES generated from the trained model and known natural product SMILES from the COCONUT database. NPClassifier is a deep learning tool that considers structural features (counted Morgan fingerprints), taxonomy of the producing organism, nature of the biosynthetic pathway, and biological activity to characterize molecules in a holistic natural product classification framework. Despite this, 7,779,787 (12%) of the generated valid SMILES received no pathway classification – a larger fraction than 35,708 (9%) of the known natural product SMILES that also received no pathway classification. It has been reported that deficiencies in NPClassifier can be traced back to limitations in its training data as the model relies on existing knowledge of natural products to classify molecules based on structural similarities. The comparatively higher percentage of generated SMILES with no NPClassifier pathway class suggests the presence of either synthetic structural features, or novel natural product class(es). However, similarities in the natural product-likeness score distributions of the generated and known datasets (KL divergence of 0.064 nats) suggests promising potential toward the latter. The remaining 59,284,417 (88%) of the generated valid natural product-like SMILES were annotated with a comparable distribution of biosynthetic pathways as known natural products from the COCONUT database with a KL divergence of 0.047 nats.

Finally, to describe physiochemical space covered by known natural products in the COCONUT database versus the >67 million natural product-like generated database, 10 physiochemical molecular descriptors for each molecule were calculated using RDkit: Number of aromatic rings; Number of aliphatic rings; Wildman-Crippen LogP (partition coefficient); Molecular weight; Number of

hydrogen bond acceptors; Number of hydrogen bond donors; Number of heteroatoms; Topological polar surface area (TPSA); Number of rotatable bonds; and Number of valence electrons.

T-distributed stochastic neighbour embedding (t-SNE) dimensionality reduction of the 10 calculated molecular descriptors into two-dimensional space was performed and plotted to visualize both physiochemical and structural space coverage.

The t-SNE 2D comparison showed a significant increase in physiochemical space covered by generated SMILES, indicating the presence of structurally novel natural product-like molecules in the generated database. Density plots showing the concentration of structures across the t-SNE 2D projected space also highlight the significantly expanded structural space offered by the generated database even in overlapping physiochemical space. Overall, this workflow has enabled the significant expansion of a database of 67,064,204 characterized natural product-like molecules, greatly increasing natural product chemical space by 165-fold over the currently estimated 400,000 natural products known.

Tandem mass spectral reference library

Using the binned collision energy specific molecule-to-MS/MS model above, one collision energy specific MS/MS spectrum is generated for each collision energy bin, for each unique charged molecule. The collision energy specific MS/MS spectra can be generated for both known molecules and generated molecules (e.g., the >67 million natural product-like compounds) to make spectra available for previously untested molecules. Thus, the library can include an experimentally generated mass spectrum (i.e., an actual mass spectrum) for each of a plurality of known molecules, thus incorporating currently available MS/MS spectral database information, and one or more mass spectra comprising peak intensities predicted using the model trained using the method of Figure

1B, for each of a plurality of natural product-like compounds.

These were compiled to yield a comprehensive high fidelity tandem mass spectral reference library

It will be appreciated that many further modifications and permutations of various aspects of the described embodiments are possible. Accordingly, the described aspects are intended to embrace all such alterations, modifications, and variations that fall within the spirit and scope of the appended claims.

Throughout this specification and the claims which follow, unless the context requires otherwise, the word "comprise", and variations such as "comprises" and "comprising", will be understood to imply the inclusion of a stated integer or step or group of integers or steps but not the exclusion of any other integer or step or group of integers or steps.

Claims

1. A method for training a model to predict a collision-energy specific tandem mass spectrum (MS/MS) of a molecule, comprising:

receiving input data comprising molecules and, for each molecule, an adduct type, a molecular formula, Simplified Molecular Input Line Entry System (SMILES) string, chemical identifier for the molecule, and ground truth intensities of a mass spectrum generated for the molecule at a specific collision energy;

generating a directed acyclic graph (DAG) comprising predicted fragments for each molecule based on the respective adduct type, SMILES string, chemical identifier, and the ground truth intensities;

training a fragmentation model based on the DAG for each molecule and the respective molecular input and ground truth intensities, to generate most probable fragments based on molecular inputs;

separating the input data into bins corresponding to non-overlapping collision energy ranges for which identical ones of the molecules have mass spectra with at least a minimum threshold similarity;

freezing the fragmentation model and generating further fragments, using the frozen fragmentation model, from input data corresponding to further molecules and respective ground truth mass spectra each generated at a specific collision energy, and training a plurality of intensity models, each intensity model being trained on the input data corresponding to a respective one of the bins, to predict a spectrum each of the further molecules, by:

predicting an intensity of a MS/MS peak for a mass of each further fragment and for a range of masses corresponding to potential hydrogen additions to, and hydrogen removals from, the further fragment; and

updating the respective intensity model to maximise weighted cosine similarity between the predicted MS/MS peak intensities and the ground truth mass spectrum.

2. The method of claim 1, comprising iteratively increasing a number of the bins until the mass spectra of the identical ones of the molecules in every bin have at least the minimum threshold similarity.

3. The method of claim 1 or 2, wherein the non-overlapping collision energy ranges are each 10 eV.

4. The method of claim 1, wherein the bins are successively generated starting from a first bin having a collision energy range including a lowest collision energy and an upper collision energy bound that is increased to a largest value for which the minimum threshold similarity is met, and each subsequent bin is then sequentially generated starting form a collision energy that is 1 eV higher than the largest collision energy for a previous said bin.

5. A model trained in accordance with the method of any one of claims 1 to 4.

6. A tandem mass spectral reference library comprising:
a plurality of molecules and, for each molecule, one or more mass spectra comprising peak intensities predicted using the model of claim 5.

7. The tandem mass spectral reference library of claim 6, wherein the plurality of molecules comprises natural product-like compounds.

8. The tandem mass spectral reference library of claim 6 or 7, wherein the plurality of molecules comprises non-natural product-like compounds.

9. The tandem mass spectral reference library of any one of claims 6 to 8, wherein the plurality of molecules comprises new molecules generated by:
tokenizing SMILES of known molecules, into a plurality of tokens;

using a molecule generation model to assemble new SMILES using the tokens; and

updating the model by comparing the new SMILES to a pre-existing set of SMILES.

10. The tandem mass spectral reference library of claim 9, wherein SMILES of known molecules are SMILES of natural products, the molecule generation model learning a relationship between the tokens of the SMILES.

11. The tandem mass spectral reference library of any one of claims 6 to 10, comprising an experimentally generated mass spectrum for each of a plurality of known molecules.